polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids

Lindsay V. Clark, Alexander E. Lipka, and Erik J. Sacks

Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

ORCID ID: 0000-0002-3881-9252 (L.V.C.), 0000-0003-1571-8528 (A.E.L.)

Software archived at: https://doi.org/10.5281/zenodo.1143744

Data and scripts for analysis archived at: https://doi.org/10.13012/B2IDB-9729830_V2

Running title: polyRAD: genotyping in polyploids

Keywords: next-generation DNA sequencing; polyploidy; Bayesian genotype calling; single nucleotide polymorphism; genotype imputation

Correspondence: Lindsay V. Clark, Department of Crop Sciences, University of Illinois at Urbana-Champaign, 1201 W. Gregory Dr., Urbana, IL 61801. Phone: +1 217-333-3420. Email: lvclark@illinois.edu

Abstract

Low or uneven read depth is a common limitation of genotyping-by-sequencing (GBS) and restriction site-associated DNA sequencing (RAD-seq), resulting in high missing data rates, heterozygotes miscalled as homozygotes, and uncertainty of allele copy number in heterozygous polyploids. Bayesian genotype calling can mitigate these issues, but previously has only been implemented in software that requires a reference genome or uses priors that may be inappropriate for the population. Here we present several novel Bayesian algorithms that estimate genotype posterior probabilities, all of which are implemented in a new R package, polyRAD. Appropriate priors can be specified for mapping populations, populations in Hardy-Weinberg equilibrium, or structured populations, and in each case can be informed by genotypes at linked markers. The polyRAD software imports read depth from several existing pipelines, and outputs continuous or discrete numerical genotypes suitable for analyses such as genome-wide association and genomic prediction.

Introduction

Approximately 70% of vascular plant species are recent polyploids, yet genomic resources and bioinformatics tools for polyploids typically lag behind those for diploids (Moghe and Shiu 2014; Renny-Byfield and Wendel 2014; Bourke *et al.* 2018b). Reduced representation DNA sequencing methods, such as genotyping-by-sequencing (GBS) and restriction site-associated DNA sequencing (RAD-seq), have made high-density genotyping considerably more accessible and affordable (Poland and Rife 2012; Davey *et al.* 2013). However, the two most popular pipelines for processing GBS and RAD-seq data, Stacks (Catchen *et al.* 2013) and TASSEL (Glaubitz *et al.* 2014), do not output polyploid genotypes. Though pipelines for polyploids are available, each have limitations that prevent their general application. For example, the UNEAK pipeline is designed for diploidized polyploids only (Lu *et al.* 2013). HaploTag is specialized for self-fertilizing polyploids (Tinker *et al.* 2016). FreeBayes and GATK can output polyploid genotypes, but require a reference genome (McKenna *et al.* 2010; Garrison and Marth 2012). The software EBG imports read depth from other pipelines to estimate auto- or allopolyploid genotypes (Blischak *et al.* 2018) but requires allele frequency estimations from the parent species for allopolyploids. The R package updog estimates polyploid genotypes from read depth, modeling preferential pairing and accounting for multiple technical issues that can arise with sequencing data, and can output posterior mean genotypes reflecting genotype uncertainty (Gerard *et al.* 2018), but requires excessive amounts of computational time to run. SuperMASSA (Serang *et al.* 2012) and fitPoly (Voorrips *et al.* 2011) were originally designed for calling polyploid genotypes from fluorescence-based SNP assays and have been adapted for sequencing data, but fail to call genotypes when low read depth results in high variance of read depth ratios. Thus, important staple crops such as wheat, potato, sweet potato, yam, and plantain are underserved by existing genotyping software, limiting our ability to perform marker-assisted selection, while yield increases from breeding are not keeping pace with projected food demands (Ray *et al.* 2013).

We present a new R package, polyRAD, for genotype estimation from read depth in polyploids and diploids. The software polyRAD is designed on the principle originally proposed by Li (2011) that it is not necessary to call genotypes with complete certainty in order to make useful inferences from sequencing data. Initially, SNP discovery is performed by other software such as TASSEL (Glaubitz *et al.* 2014) or Stacks (Catchen *et al.* 2013), with or without a reference genome, then allelic read depth is imported into polyRAD from those pipelines or the read counting software TagDigger (Clark and Sacks 2016). In polyRAD, one or several ploidies can be specified, including any level of auto- and/or allopolyploidy, allowing inheritance modes to vary across the genome. Genotype probabilities are estimated by polyRAD under a Bayesian framework, where priors are based on mapping population design, Hardy-Weinberg equilibrium (HWE), or population structure, with or without linkage disequilibrium (LD) and/or self-fertilization. Multi-allelic loci (haplotypes) are allowed, and are in fact encouraged because LD within the span of one RAD tag is not informative for genotype imputation. In addition to exporting the most probable genotype for each individual and locus, continuous numerical genotypes can be exported reflecting the relative probabilities of all possible allele copy numbers, and can then be used for genome-wide association or genomic prediction in software such as GAPIT (Lipka *et al.* 2012), FarmCPU (Liu *et al.* 2016b), TASSEL (Bradbury *et al.* 2007), or rrBLUP (Endelman 2011). Discrete genotypes can also be exported for polymapR (Bourke *et al.* 2018a). polyRAD is the first Bayesian genotype caller to incorporate population structure and multiple inheritance modes, as well as the first with an option for mapping population designs other than F1 and F2. It is available at https://github.com/lvclark/polyRAD and https://CRAN.R-project.org/package=polyRAD.

Methods

Overview

polyRAD implements Bayesian genotype estimation, similar to that proposed and implemented by several other groups (Li 2011; Nielsen *et al.* 2011; Garrison and Marth 2012; Korneliussen *et al.* 2014; Maruki and Lynch 2017; Gerard *et al.* 2018; Blischak *et al.* 2018). In all polyRAD pipelines, genotype prior probabilities ($P(G_i)$) represent, for a given allele and individual, the probability that i is the true allele copy number, before taking allelic read depth into account. Genotype prior probabilities are specified from population parameters, and optionally from genotypes at linked markers (see Supplementary Methods).

For a given individual and locus, consider every sequencing read to be a Bernoulli trial, where the read either matches a given allele (success) or some other allele (failure). The probability of success is:

Eqn. 1: $\pi_i = (1 - c) * \frac{i}{k} + c * p$,

where c is the cross-contamination rate, i is the allele copy number in the genotype, k is the ploidy, and p is the allele frequency in the population. The c parameter is important for identifying homozygotes that could otherwise be misidentified as heterozygotes. For GBS and RAD-seq data, c is estimated by including a negative control in library preparation, i.e. of the set of ligation reactions with barcoded adapters, one that has no genomic DNA added. The sequence read depth for this blank barcode is then divided by the mean read depth of non-blank barcodes in order to estimate c. Our model assumes c to be constant across loci, under the assumption that most errors are due to contamination during library preparation. In practice we have found *c* to typically be 1/1000 (unpublished data), and expect it to be more substantial than errors arising from the sequencing technology, which will tend to produce haplotypes not found elsewhere in the data set. Therefore, although it is known that sequencing error can vary from locus to locus depending on sequence context (Nakamura *et al.* 2011), polyRAD does not estimate sequence error on a per-locus basis. Rare loci with very high sequencing error rates may exhibit underestimated likelihoods of homozygosity.

Gerard et al. (2018) observed overdispersion in the distribution of sequence read depth, indicating that in reality π_i varies from sample to sample. We have observed the same in our datasets, likely due to factors such as differing contamination rates among samples, restriction cut site variation, and differences in size selection among libraries. Therefore, following Gerard et al. (2018), we model allelic read depth as following a beta-binomial distribution rather than a binomial distribution. For every possible allele copy number at a given locus and individual, the following equation is used to estimate the likelihood of the observed read depth using the beta-binomial probability mass function:

Eqn. 2: $$L(a,b|G_i) = \binom{a+b}{a} * \frac{B[d*\pi_i+a,\ d*(1-\pi_i)+b]}{B[d*\pi_i,\ d*(1-\pi_i)]},$$

where *a* is the number of reads for a given allele at a given locus, *b* is the number of reads for other alleles at that locus, G_i is the state in which a locus has *i* copies of a given allele, *B* is the beta function, and *d* is the overdispersion parameter. The parameter *d* is set to nine by default given our observations of overdispersion in empirical data, and can be increased to model less overdispersion and vice versa. The function *TestOverdispersion* is included in polyRAD to assist the user in determining the optimal value of *d*. Although overdispersion is likely to vary from locus to locus, polyRAD uses a single estimate in order to save computational time. The lower *d* is, the more influence genotype prior probabilities have on genotype estimates.

From the priors and likelihoods, a posterior probability can then be estimated for each possible allele copy number for each individual and allele using Bayes' theorem (Shiryaev 2011):

Eqn. 3: $P(G_i|a,b) = \frac{L(a,b|G_i)*P(G_i)}{\sum_{i=0}^{k} L(a,b|G_i)*P(G_i)}$,

where all terms are as previously described.

Bayesian genotype estimation allows correction of genotyping errors in diploids and polyploids, i.e. when an individual is truly heterozygous but only one allele was sequenced, or when an individual appears heterozygous due to sequencing error or contamination but is truly homozygous. It also enables estimation of allele dosage in heterozygous polyploid genotypes. Moreover, genotype posterior probabilities are more influenced by priors when read depth is low, and by genotype likelihoods derived from allelic read depth when read depth is high. When read depth is zero for a given individual and locus, genotype posterior probabilities are equal to priors, and thus missing and non-missing data are handled within one coherent paradigm. It is therefore not necessary to impute missing genotypes in a second step if the priors are sufficiently informative.

For export to other software, as well as iteration within the polyRAD pipelines, a given allele's posterior mean genotype (*pmg*) is a mean of the number of copies of that allele, with the posterior genotype probabilities (Eqn. 3) serving as weights, as in Guan and Stephens (2008). Thus, for an individual and allele, *pmg* is calculated as:

Eqn. 4: $pmg = \sum_{i=0}^{k} P(G_i|a,b) * i$, where all terms are as previously described. Additional details and equations for specification of prior genotype probabilities and estimation of other parameters are provided in Supplementary Materials. A flow chart of how this Bayesian genotypic estimation is implemented into polyRAD is displayed in Fig. 1. In brief, for mapping populations, genotype priors are specified based on parental genotypes and progeny allele frequencies, and all parameters are estimated once. For diversity panels, genotype priors are adjusted and parameters re-estimated iteratively until allele frequencies converge. Source code is available at https://github.com/lvclark/polyRAD, archived at Zenodo (doi: 10.5281/zenodo.1143744).

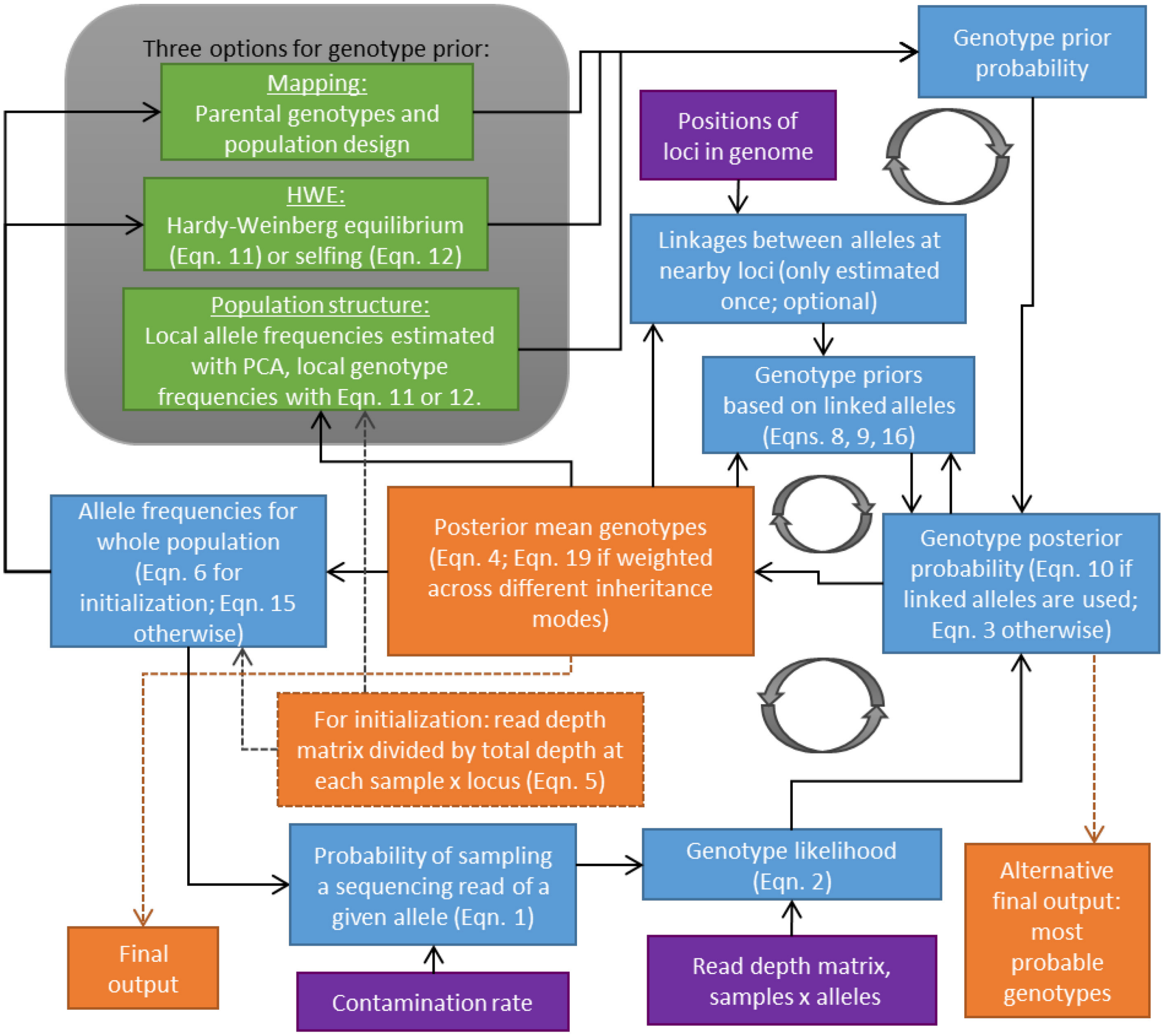

Three options for genotype prior:
Mapping:
Parental genotypes and population design
HWE:
Hardy-Weinberg equilibrium (Eqn. 11) or selfing (Eqn. 12)
Population structure:
Local allele frequencies estimated with PCA, local genotype frequencies with Eqn. 11 or 12.
Genotype prior probability
Positions of loci in genome
Linkages between alleles at nearby loci (only estimated once; optional)
Genotype priors based on linked alleles (Eqns. 8, 9, 16)
Allele frequencies for whole population (Eqn. 6 for initialization; Eqn. 15 otherwise)
Posterior mean genotypes (Eqn. 4; Eqn. 19 if weighted across different inheritance modes)
Genotype posterior probability (Eqn. 10 if linked alleles are used; Eqn. 3 otherwise)
For initialization: read depth matrix divided by total depth at each sample x locus (Eqn. 5)
Probability of sampling a sequencing read of a given allele (Eqn. 1)
Genotype likelihood (Eqn. 2)
Alternative final output: most probable genotypes
Final output
Contamination rate
Read depth matrix, samples x alleles

Fig. 1. Overview of polyRAD algorithms for genotype estimation. Genotype posterior probabilities are estimated iteratively until allele frequencies converge, except in the case of mapping populations, where allele frequencies are only estimated once. Purple boxes indicate inputs to the pipeline (read depth, contamination rate, and optionally, genomic positions of loci). Blue boxes indicate estimated parameters (allele frequencies, genotype likelihoods and prior and posterior probabilities, linkage between alleles, and probability of sampling each allele). Green boxes indicate alternative methodologies for genotype prior probability estimation (mapping, HWE, and population structure). Priors for the HWE and population structure models can be adjusted for self-fertilization according to de Silva et al. (2005). Orange boxes indicate sample × allele matrices indicating approximate allele copy number. Dashed arrows indicate steps that happen only once at the beginning or end of the pipeline, whereas solid arrows indicate

iterative steps. Circular arrows highlight cycles of iteration. Eqns. 1-4 are provided in the main manuscript, and Eqns. 5-19 are provided in Supplemental Materials.

Example use

Executable examples are provided in the vignette and manual distributed with polyRAD. Here we provide an additional brief example. Box 1 illustrates the use of polyRAD on a diversity panel of a generic tetraploid species with a reference genome. Tools from the Bioconductor package VariantAnnotation (Obenchain *et al.* 2014) are used by the polyRAD function *VCF2RADdata* for import of a VCF file to the polyRAD-specific "RADdata" format. SNP filtering criteria are specified with the *min.ind.with.reads* and *min.ind.with.minor.allele* arguments to indicate the minimum number of individuals that must have more than zero reads of a locus, and the minimum number of individuals that must have reads of the minor allele, respectively. The *possiblePloidies* argument indicates that the inheritance mode could be allotetraploid (*c(2,2)*) or autotetraploid (*4*). Any ploidy may be specified with *possiblePloidies*, for example *8* for auto-octoploid, with the only limitation that all subgenomes in an allopolyploid must have the same ploidy. By default, *VCF2RADdata* groups SNP alleles into haplotypes that appear to have come from the same RAD tag, the size of which is specified by *tagsize*, in basepairs. Negative controls are indicated with *SetBlankTaxa*, and the contamination rate is estimated with *EstimateContaminationRate*. The function *IteratePopStructLD* is then used for genotype estimation, taking both population structure and LD into account. The probabilistic principal components analysis method from the Bioconductor package pcaMethods (Stacklies *et al.* 2007) is used internally by *IteratePopStructLD* in order to estimate population structure. The *LDdist* argument indicates the distance in basepairs within which to search for alleles at other loci that can help predict copy number of a given allele. Once genotype posterior probabilities are estimated, other parameters are cleared from memory using the *StripDown* function. Continuous numerical genotypes are then formatted for GAPIT (Lipka *et al.* 2012) using the

ExportGAPIT function. Alternative functions are listed in Table 1. A very similar script could be used for a species without a reference genome, with *IteratePopStruct* in place of *IteratePopStructLD*, and a different import function for the appropriate non-reference pipeline.

```
library(polyRAD)
library(VariantAnnotation)
# prepare the VCF file for import
myvcf <- "somegenotypes.vcf"
myvcfbg <- bgzip(myvcf)
indexTabix(myvcfbz, format = "vcf")
# import VCF into a RADdata object
myRAD <- VCF2RADdata(myvcfbg,
                     tagsize = 64,
                     min.ind.with.reads = 300,
                     min.ind.with.minor.allele = 15,
                     possiblePloidies = list(c(2,2), 4))
# estimate contamination rate
myRAD <- SetBlankTaxa(myRAD, c("blank1", "blank2"))
myRAD <- EstimateContaminationRate(myRAD)
# genotype estimation with pop. structure pipeline
myRAD <- IteratePopStructLD(myRAD, LDdist = 5e4)
# free up memory
myRAD <- StripDown(myRAD)
# export for GAPIT
myGM_GD <- ExportGAPIT(myRAD)
```

Box 1. Example R script using polyRAD. Read depth is imported from a VCF file, genotypes are estimated using population structure and LD, and continuous numerical genotypes are formatted for GAPIT.

Table 1. Overview of main polyRAD functions.

Import functions	
VCF2RADdata	Imports any VCF with an allelic read depth (AD) field, such as those exported by TASSEL-GBSv2 or GATK.
readTagDigger	Imports CSV file of read depth output by TagDigger.
readStacks	Reads catalog and matches files from Stacks.
readTASSELGBSv2	Reads SAM and TagTaxaDist files from TASSEL-GBSv2.
readHMC	Reads files output by UNEAK.
Genotype estimation functions	
PipelineMapping2Parents	For mapping populations with any number of generations of backcrossing, intermating, and/or selfing.
IterateHWE	For diversity panels without population structure.[a]
IterateHWE_LD	For diversity panels with LD and without population structure.[a]
IteratePopStruct	For diversity panels with population structure.[a]
IteratePopStructLD	For diversity panels with population structure and LD.[a]
Export functions	
ExportGAPIT	Format genotypes for the *GD* and *GM* arguments of GAPIT or FarmCPU.
Export_rrBLUP_Amat	Format genotypes for the *A.mat* function in rrBLUP.
Export_rrBLUP_GWAS	Format genotypes for the *GWAS* function in rrBLUP.
Export_TASSEL_Numeric	Write file formatted for TASSEL with continuous numeric genotypes.
Export_polymapR	Format genotypes for the polymapR package.
GetWeightedMeanGenotypes	Create a matrix of continuous numeric genotypes.
GetProbableGenotypes	Create a matrix of discrete genotypes, indicating the most probable genotype for each individual and allele.

[a] The rate of self-fertilization can be specified for self-compatible plant species.

Testing

To test the accuracy of polyRAD, we used datasets from three previously studied populations: 1) RAD-seq data and GoldenGate SNP genotypes from a diversity panel (n = 565) of the outcrossing, diploidized allotetraploid grass *Miscanthus sinensis* (Clark *et al.* 2014), 2) RAD-seq data and GoldenGate SNP genotypes from a bi-parental F_1 mapping population (n = 275) of *M. sinensis* (Liu *et al.* 2016a), and 3) SNP array genotypes from a biparental F_1 mapping population of autotetraploid potato (n = 238) (da Silva *et al.* 2017). Allelic read depth at simulated RAD-seq markers was generated from the GoldenGate or SNP array genotypes, with overall locus depth drawn from a gamma distribution to resemble depth of actual RAD-seq markers (shape = 2 and scale = 5). The read depth for an individual genotype was also sampled from a gamma distribution, with the shape equal to the locus depth divided by 10, and scale = 10. The read depth for each allele was then sampled from the beta-binomial distribution as described in Eqn. 2, with $d = 9$ and $c = 0.001$. The *M. sinensis* diversity panel included 395 GoldenGate markers, plus real RAD-seq data for those same individuals across 3290 tag locations within 20 kb of any GoldenGate markers, called with the TASSEL GBS v2 pipeline (Glaubitz *et al.* 2014) using the *M. sinensis* v7.1 reference genome (DOE-JGI, http://phytozome.jgi.doe.gov/). Additionally, to test the effect of ploidy within the *M. sinensis* diversity panel, tetraploidy was simulated by summing GoldenGate genotypes and RAD-seq read depth of each individual with the individual with the most similar read depth to it out of the ten individuals most closely related to it. The *M. sinensis* mapping population included 241 GoldenGate markers genotyped across 83 individuals, plus 3062 RAD-seq markers called with the UNEAK pipeline (Lu *et al.* 2013) across those 83 individuals plus an additional 192 individuals. The potato mapping population included genotypes at 2538 markers. Additional simulations using data from diversity panels of soybean (Song *et al.* 2015), apple (Chagné *et al.* 2012), and potato (Hamilton *et al.* 2011) are presented in Figs. S1-S4. In each population, the simulated and real RAD-seq data were used for genotype calling with polyRAD, EBG (Blischak *et al.* 2018), updog (Gerard *et al.* 2018), and fitPoly (Voorrips *et al.* 2011), and missing genotypes from the EBG output were imputed with LinkImpute (Money *et al.* 2015) and/or rrBLUP (Endelman 2011) as appropriate. To estimate the accuracy of genotype calling and imputation, the root mean squared error (RMSE) was calculated between numeric genotypes (ranging from zero to the ploidy) at each simulated RAD-seq marker and at the GoldenGate or SNP array marker from which it was derived.

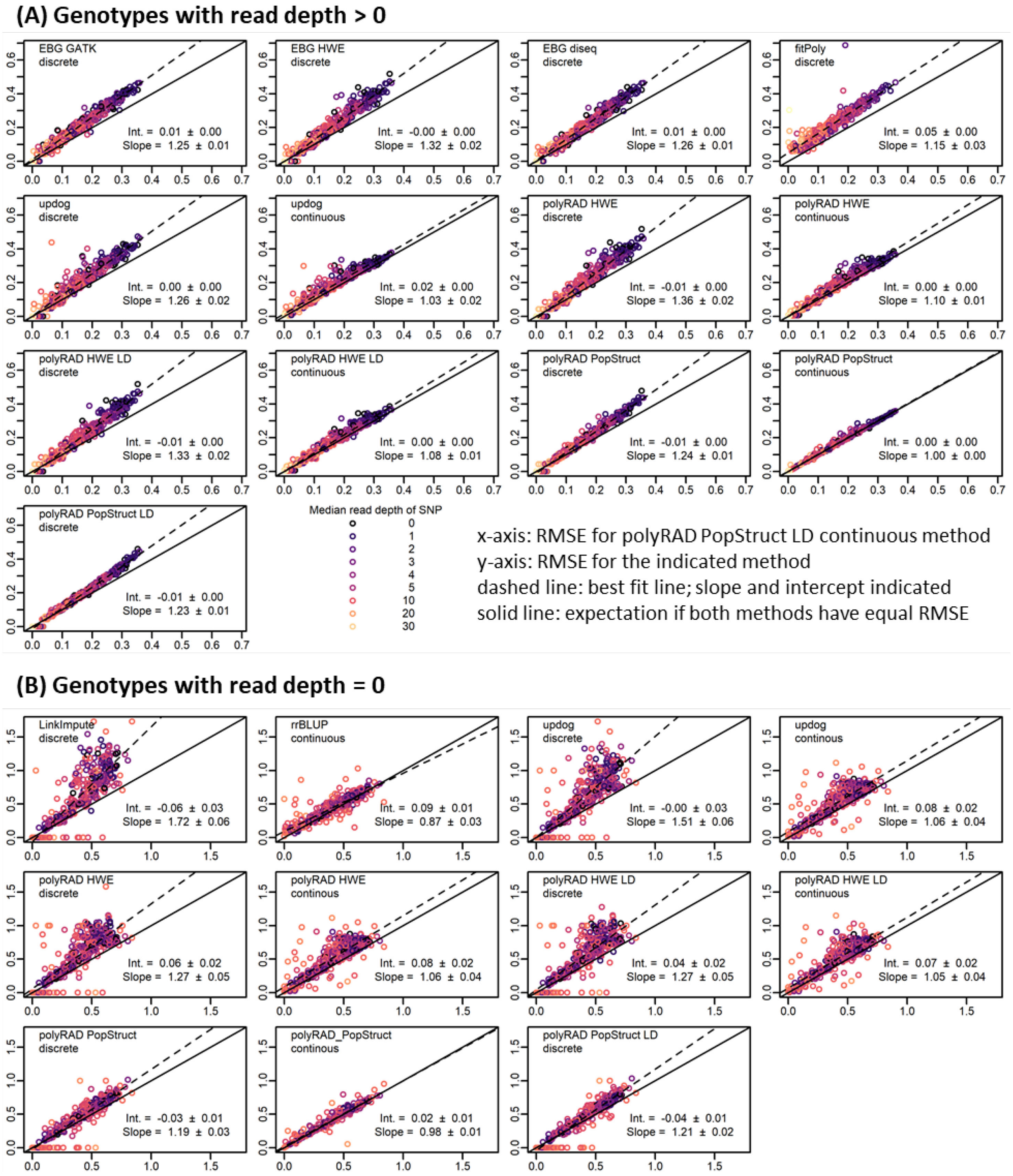
Data Availability

Data and scripts for analysis are available at https://doi.org/10.13012/B2IDB-9729830_V2. Supplementary text, equations, and figures have been deposited at Figshare: https://doi.org/10.25387/g3.7370999 (https://figshare.com/s/f7fe2995eacbfd7e6066).

Results and discussion

Accuracy of polyRAD

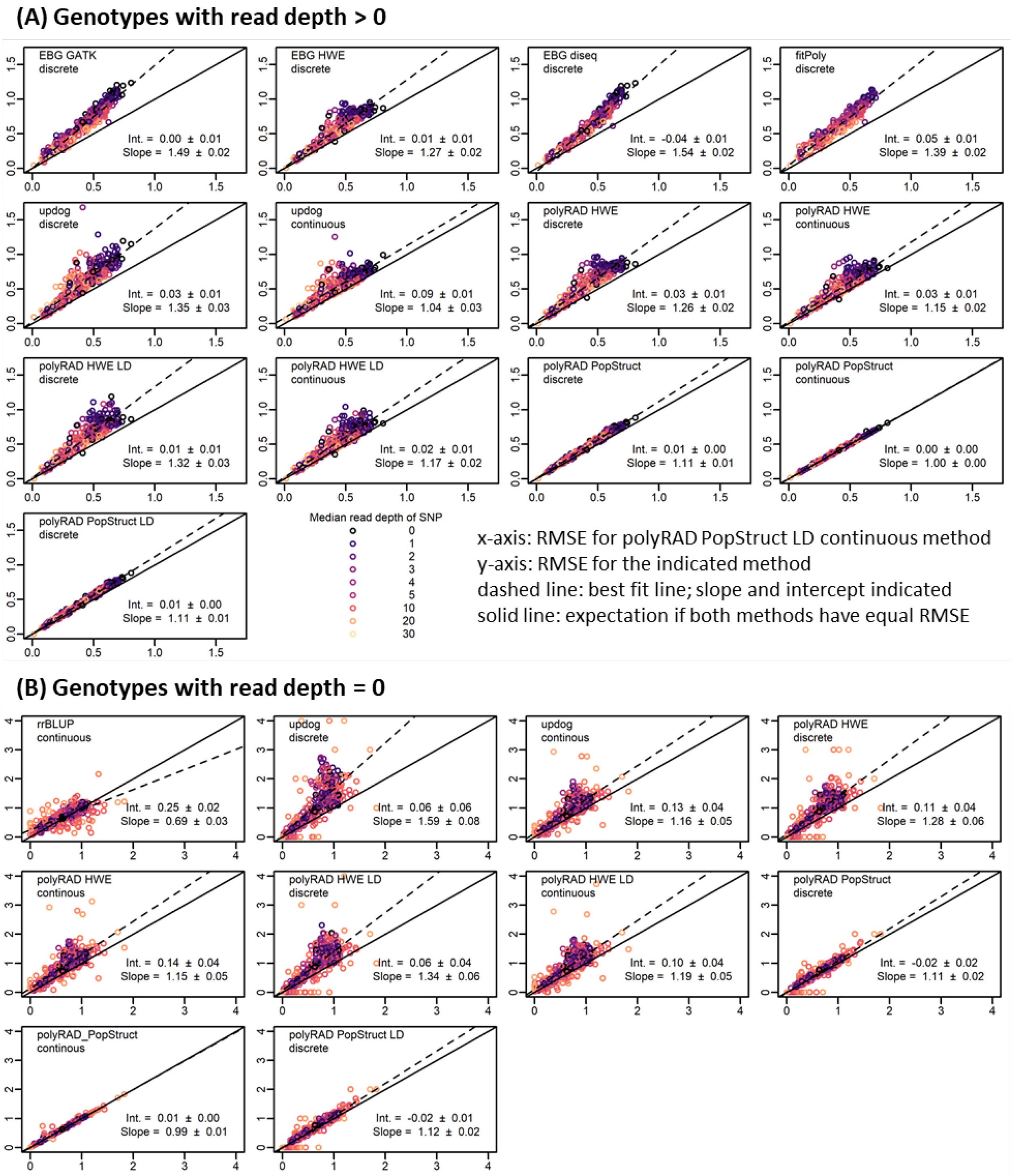
In the *M. sinensis* diversity panel, polyRAD showed improved genotype accuracy over the HWE, disequilibrium, and GATK methods implemented in EBG, as well as fitPoly, particularly at low read depths (Figs. 2A and 3A). polyRAD also showed a modest improvement in accuracy across all read depths as compared to updog (Figs. 2A and 3A) while needing approximately two orders of magnitude less processing time than updog. Under the HWE model in polyRAD with discrete genotypes output, errors in genotypes with more than zero reads were similar to those from the HWE model of EBG in both diploid and tetraploid systems (Figs. 2A and 3A). However, when priors in polyRAD were based on population structure, errors decreased, particularly in tetraploids and at low read depth (Figs. 2A and 3A). In diploids and tetraploids respectively using the polyRAD population structure model with discrete genotypes, error (RMSE) was reduced by 14.6% (SE 1.0%) and 23.5% (SE 0.6%) relative to the GATK model, by 10.5% (SE 0.9%) and 11.8% (SE 0.5%) relative to the EBG HWE model, by 26.0% (SE 1.2%) and 25.6% (SE 0.6%) relative fitPoly, and by 8.0% (SE 1.0%) and 18.0% (SE 0.7%) relative to discrete genotype output by the updog "norm" model. Given the known population structure in *M. sinensis* (Clark *et al.* 2014), it is unsurprising that a population structure-aware genotyping method would be more accurate than those based on HWE or otherwise not accounting for population structure. For genotypes with zero reads, imputation was most accurate when it accounted for population structure, using either polyRAD or rrBLUP (Fig. 2B and 3B). Although modeling LD did not improve accuracy in *M. sinensis* (Figs. 2 and 3), likely due to low LD as a result of outcrossing (Slavov *et al.* 2014), modeling LD did improve accuracy in wild soybean, apple, and a simulated inbreeding allohexaploid (Figs. S1, S2, and S3, and Supporting Results). In a diversity panel of tetraploid potato, accuracy was improved by modeling population structure but not LD (Fig. S4 and Supporting Results).



x-axis: RMSE for polyRAD PopStruct LD continuous method
y-axis: RMSE for the indicated method
dashed line: best fit line; slope and intercept indicated
solid line: expectation if both methods have equal RMSE

Fig. 2. Genotyping error of EBG, fitPoly, updog, polyRAD, LinkImpute, and rrBLUP in a diversity panel of 565 diploid *Miscanthus sinensis*. The benefits of incorporating population structure into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a

scale of 0 to 2. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 395 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The “norm” model was used with updog. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.


(A) Genotypes with read depth > 0

x-axis: RMSE for polyRAD PopStruct LD continuous method
y-axis: RMSE for the indicated method
dashed line: best fit line; slope and intercept indicated
solid line: expectation if both methods have equal RMSE

(B) Genotypes with read depth = 0

Fig. 3. Genotyping error of EBG, fitPoly, updog, polyRAD, and rrBLUP in a simulated tetraploid diversity panel derived from genotypes of 565 diploid *Miscanthus sinensis*. The benefits of incorporating population structure into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 4. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 395 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The “norm” model was used with updog. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type. LinkImpute was not included given that it works for diploids only.

In diploid *M. sinensis* and tetraploid potato F1 mapping populations, polyRAD outperformed the GATK method, fitPoly, and updog, particularly when linked markers were used for informing the priors in polyRAD (Figs. 4A and 5A). In diploids and tetraploids respectively using genotypes with non-zero read depth, error (RMSE) using the polyRAD linkage model with discrete genotypes was reduced by 31.6% (SE 2.2%) and 48.0% (SE 0.4%) with respect to the GATK model, and 1.5% (SE 3.1%) and 17.1% (SE 0.6%) with respect to the updog "f1" model with discrete genotypes. For diploids, error was reduced by 39.8% (SE 2.5%) using polyRAD with respect to fitPoly, while for tetraploids fitPoly failed for all markers. For imputation, polyRAD using the linkage model performed similarly to LinkImpute and rrBLUP (Figs. 4B and 5B). Although only F1 populations are presented here, many other population types are supported in polyRAD.

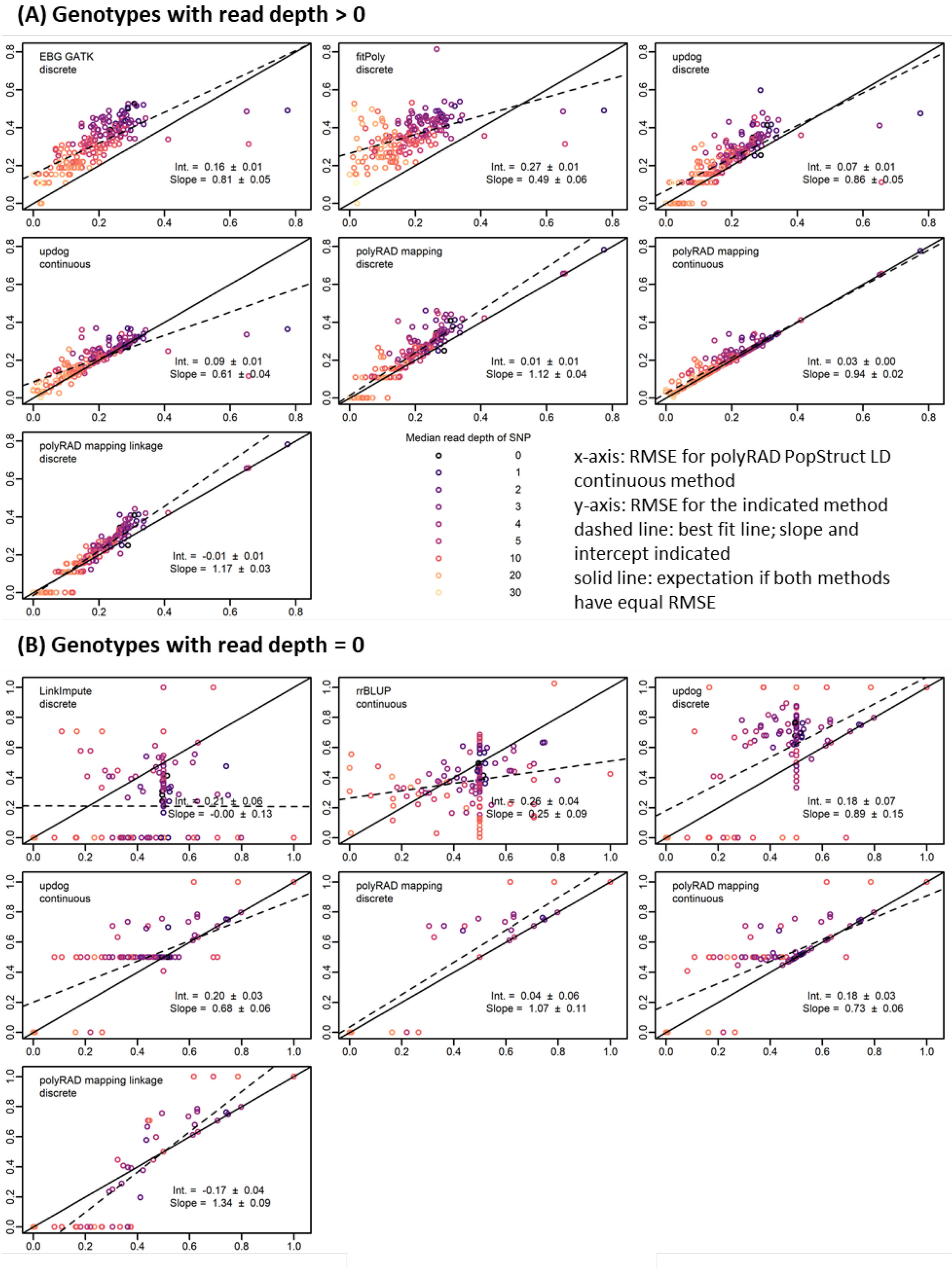
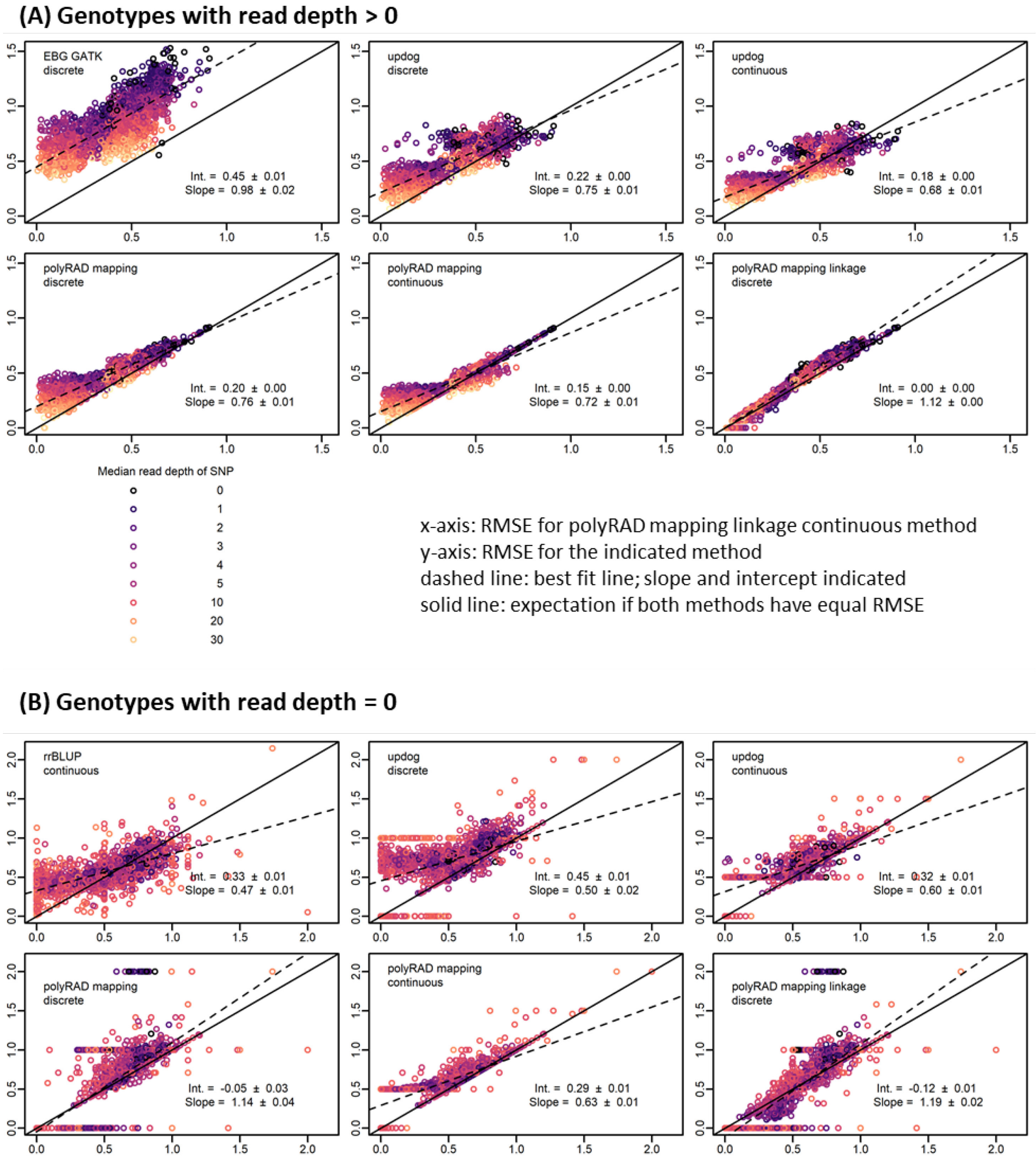

(A) Genotypes with read depth > 0
EBG GATK discrete
Int. = 0.16 ± 0.01
Slope = 0.81 ± 0.05
fitPoly discrete
Int. = 0.27 ± 0.01
Slope = 0.49 ± 0.06
updog discrete
Int. = 0.07 ± 0.01
Slope = 0.86 ± 0.05
updog continuous
Int. = 0.09 ± 0.01
Slope = 0.61 ± 0.04
polyRAD mapping discrete
Int. = 0.01 ± 0.01
Slope = 1.12 ± 0.04
polyRAD mapping continuous
Int. = 0.03 ± 0.00
Slope = 0.94 ± 0.02
polyRAD mapping linkage discrete
Int. = -0.01 ± 0.01
Slope = 1.17 ± 0.03
0.0
0.2
0.4
0.6
0.8
Median read depth of SNP
0
1
2
3
4
5
10
20
30
x-axis: RMSE for polyRAD PopStruct LD continuous method
y-axis: RMSE for the indicated method
dashed line: best fit line; slope and intercept indicated
solid line: expectation if both methods have equal RMSE
(B) Genotypes with read depth = 0
LinkImpute discrete
Int. = 0.21 ± 0.06
Slope = -0.00 ± 0.13
rrBLUP continuous
Int. = 0.26 ± 0.04
Slope = 0.25 ± 0.09
updog discrete
Int. = 0.18 ± 0.07
Slope = 0.89 ± 0.15
updog continuous
Int. = 0.20 ± 0.03
Slope = 0.68 ± 0.06
polyRAD mapping discrete
Int. = 0.04 ± 0.06
Slope = 1.07 ± 0.11
polyRAD mapping continuous
Int. = 0.18 ± 0.03
Slope = 0.73 ± 0.06
polyRAD mapping linkage discrete
Int. = -0.17 ± 0.04
Slope = 1.34 ± 0.09
1.0

Fig. 4. Genotyping error of EBG, fitPoly, updog, polyRAD, LinkImpute, and rrBLUP in an F1 mapping population of 83 diploid *Miscanthus sinensis*. The benefits of incorporating linkage into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 2. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 241 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The “f1” model was used with updog. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.



x-axis: RMSE for polyRAD mapping linkage continuous method
y-axis: RMSE for the indicated method
dashed line: best fit line; slope and intercept indicated
solid line: expectation if both methods have equal RMSE

Fig. 5. Genotyping error of EBG, updog, polyRAD, and rrBLUP in an F1 mapping population of tetraploid potato with 238 progeny. The benefits of incorporating linkage into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 4. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 2538 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD mapping method with linkage is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The "f1" model was used with updog. fitPoly results are omitted since it failed for all markers, and LinkImpute was not run since LinkImpute is for diploids only. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.

Genotyping error was also reduced 10-15% in most cases by exporting genotypes as continuous numerical variables (posterior mean genotypes), rather than discrete values (Figs. 2-5). For example, in a diploid, a true heterozygote (numeric value of 1) with reads only for the reference allele might erroneously be called as zero (homozygous for the reference allele) if only the most probable genotype is exported. However, the genotype could be called 0.4 if continuous genotypes are allowed, indicating that there is a 60% chance of it being a homozygote and 40% chance of it being a heterozygote, and thereby reducing the error from 1.0 to 0.6. Similarly in polyploids, continuous numerical genotypes can correct for errors in allele copy number estimation of heterozygotes.

Downstream applications and implications for sequencing strategies

The genotyping methods implemented in polyRAD will have the most benefit for marker analysis where 1) the accuracy of individual genotypes is important, and 2) genotypes can be treated as continuous rather than discrete variables. The use of continuous versus discrete genotypes has been demonstrated to increase power for genome-wide association studies (GWAS) (Grandke *et al.* 2016) and genomic prediction (Oliveira *et al.* 2018) in polyploids. More generally, we anticipate that analyses that seek to quantify marker-trait associations in a population of individuals, including GWAS, quantitative trait locus mapping, and genomic prediction methods involving variable selection, will especially benefit from polyRAD. By reducing genotyping error, polyRAD will increase the power of these methods to detect true associations. Analyses that will benefit less from polyRAD genotyping are those where an average is taken across many genotypes in order to estimate a statistic, such as allele frequencies in a population or overall relatedness of individuals (including kinship-based methods of genomic prediction), because genotyping errors generally are not biased towards one allele or the other and tend to balance out over many individuals and loci (Buerkle and Gompert 2013; Dodds *et al.* 2015).

The advantages of polyRAD for accurate genotyping at low sequence read depth alter the economics of sequence-based genotyping, enabling researchers to purchase fewer sequencing lanes, multiplex more samples per lane, and/or retain more markers during filtering. In particular, for protocols using restriction enzymes where read depth varies considerably from locus to locus depending on fragment size (Beissinger *et al.* 2013; Davey *et al.* 2013; Andrews *et al.* 2016), there are diminishing returns on increasing the per-sample read depth, because some loci receive far more reads than are needed for accurate genotyping while other loci remain poor quality. Using population structure and linkage between loci, polyRAD uses information from high-depth markers to improve genotyping accuracy of low-depth markers, helping to maximize the useful information that is obtained from sequencing data. This advance is expected to improve breeding efficiency and economics.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1661490. The authors thank Marcus Olatoye, Per McCord, Joyce Njuguna, Wittney Mays, and Jiale He for testing the polyRAD software, and Associate Editor Daniel Runcie and two anonymous reviewers for comments on earlier versions of this manuscript.

Literature Cited

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe, 2016 Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17: 81–92.

Beissinger, T. M., C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson *et al.*, 2013 Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193: 1073–1081.

Blischak, P. D., L. S. Kubatko, and A. D. Wolfe, 2018 SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34: 407–415.

Bourke, P. M., G. van Geest, R. E. Voorrips, J. Jansen, T. Kranenburg *et al.*, 2018a polymapR—linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics 34: 3496-3502.

Bourke, P. M., R. E. Voorrips, R. G. F. Visser, and C. Maliepaard, 2018b Tools for Genetic Studies in Experimental Populations of Polyploids. Front. Plant Sci. 9: 513.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Buerkle, C. A., and Z. Gompert, 2013 Population genomics based on low coverage sequencing: How low should we go? Mol. Ecol. 22: 3028–3035.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013 Stacks: an analysis tool set for population genomics. Mol. Ecol. 22: 3124–40.

Chagné, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore *et al.*, 2012 Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple (M. Bendahmane, Ed.). PLoS One 7: e31745.

Clark, L. V, J. E. Brummer, K. Głowacka, M. C. Hall, K. Heo *et al.*, 2014 A footprint of past climate change on the diversity and population structure of *Miscanthus sinensis*. Ann. Bot. 114: 97–107.

Clark, L. V, and E. J. Sacks, 2016 TagDigger: user-friendly extraction of read counts from GBS and RAD-seq data. Source Code Biol. Med. 11: 11.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi *et al.*, 2013 Special features of RAD Sequencing data: implications for genotyping. Mol. Ecol. 22: 3151–3164.

Dodds, K. G., J. C. McEwan, R. Brauning, R. M. Anderson, T. C. van Stijn *et al.*, 2015 Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics 16: 1047.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome J. 4: 250–255.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv 1207.3907.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping Polyploids from Messy Sequencing Data. Genetics 210: 789–807.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9: e90346.

Grandke, F., P. Singh, H. C. M. Heuven, J. R. de Haan, and D. Metzler, 2016 Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: A comparative study in hexaploid chrysanthemum. BMC Genomics 17: 672.

Guan, Y., and M. Stephens, 2008 Practical issues in imputation-based association mapping.

PLoS Genet. 4: e1000279.

Hamilton, J. P., C. N. Hansey, B. R. Whitty, K. Stoffel, A. N. Massa *et al.*, 2011 Single nucleotide polymorphism discovery in elite north American potato germplasm. BMC Genomics 12: 302.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15: 356.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Liu, S., L. V. Clark, K. Swaminathan, J. M. Gifford, J. A. Juvik *et al.*, 2016a High density genetic map of *Miscanthus sinensis* reveals inheritance of zebra stripe. GCB Bioenergy 8: 616–630.

Liu, X., M. Huang, B. Fan, E. S. Buckler, and Z. Zhang, 2016b Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies (J. Listgarten, Ed.). PLOS Genet. 12: e1005767.

Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney *et al.*, 2013 Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9: e1003215.

Maruki, T., and M. Lynch, 2017 Genotype Calling from Population-Genomic Sequencing Data.

G3 7: 1393–1404.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Moghe, G. D., and S. H. Shiu, 2014 The causes and molecular consequences of polyploidy in flowering plants. Ann. N. Y. Acad. Sci. 1320: 16–34.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong *et al.*, 2015 LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 5: 2383–90.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa *et al.*, 2011 Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 39: e90.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–451.

Obenchain, V., M. Lawrence, V. Carey, S. Gogarten, P. Shannon *et al.*, 2014 VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. Bioinformatics 30: 2076–2078.

Oliveira, I. de B., M. F. Resende, F. Ferrao, R. Amadeu, J. Endelman *et al.*, 2018 Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. bioRxiv 432179.

Poland, J. A., and T. W. Rife, 2012 Genotyping-by-sequencing for plant breeding and genetics. Plant Genome J. 5: 92–102.

Ray, D. K., N. D. Mueller, P. C. West, and J. A. Foley, 2013 Yield Trends Are Insufficient to

Double Global Crop Production by 2050. PLoS One 8: e66428.

Renny-Byfield, S., and J. F. Wendel, 2014 Doubling down on genomes: Polyploidy and crop plants. Am. J. Bot. 101: 1711–1725.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One 7: e30906.

Shiryaev, A. N., 2011 Bayes formula. Encycl. Math. Available at: https://www.encyclopediaofmath.org//index.php?title=Bayes_formula&oldid=16075

da Silva, W., J. Ingram, C. A. Hackett, J. J. Coombs, D. Douches *et al.*, 2017 Mapping Loci That Control Tuber and Foliar Symptoms Caused by PVY in Autotetraploid Potato (*Solanum tuberosum* L.). G3 7: 3587–3595.

De Silva, H. N., A. J. Hall, E. Rikkerink, M. A. McNeilage, and L. G. Fraser, 2005 Estimation of allele frequencies in polyploids under certain patterns of inheritance. Heredity (Edinb). 95: 327–334.

Slavov, G. T., R. Nipper, P. Robson, K. Farrar, G. G. Allison *et al.*, 2014 Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. New Phytol. 201: 1227–1239.

Song, Q., D. L. Hyten, G. Jia, C. V. Quigley, E. W. Fickus *et al.*, 2015 Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 5: 1999–2006.

Stacklies, W., H. Redestig, M. Scholz, D. Walther, and J. Selbig, 2007 pcaMethods - A bioconductor package providing PCA methods for incomplete data. Bioinformatics 23: 1164–1167.

Tinker, N. A., W. A. Bekele, and J. Hattori, 2016 Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. G3 6: 857–863.

Voorrips, R. E., G. Gort, and B. Vosman, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12: 172.